Using Composite Likelihood in Loss Tomography

Weiping Zhu *member, IEEE*,

***Abstract*—Full likelihood has been widely used in loss tomography because most believe it can produce accurate estimates although the full likelihood estimators proposed so far are complex in structure and expensive in execution. We in this paper advocate a different likelihood called composite likelihood to replace the full likelihood in loss tomography for simplicity and accuracy. Using the proposed likelihood, we propose a number of explicit estimators with statistical analysis. The analysis shows all of the explicit estimators perform almost as good as the full likelihood one in a number of aspects, including asymptotic variance, computational complexity, and robustness. Although the discussion is restricted to the tree topology, the methodology proposed here is also applicable to a network of a general topology.**

***Index Terms*—Composite Likelihood, Loss tomography, Model, Maximum Likelihood, Explicit Estimator.**

I. Introduction

Loss tomography has received considerable attention in the last 10 years since it provides a new methodology to measure the internal characteristics of a large network. In contrast to direct measurement, loss tomography uses statistical inference to estimate the link-level loss rates of a network from end-to-end observations. Because of this, a number of estimators have been proposed during this period. Some of the proposed estimators have been proved to be the maximum likelihood ones [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. Unfortunately, none of the maximum likelihood estimators is presented in a closed form and only a few of them are presented with statistical properties, e.g. unbiasedness, variance and robustness. Without the properties, it is hard to compare and evaluate the proposed estimators. To assist the development of loss tomography, an analysis of a frequently used estimator is presented here that unveils the statistical logic used by the estimator and sheds light on developing simple and accurate estimators.

The estimator proposed in [1] is one of the widely used estimators in loss tomography that is designated to the tree topology and has a likelihood equation in the form of a single variable polynomial. The likelihood equation connects the pass rate of a path, from the root to an internal node, to the pass rates of the pathes from the root to the children of the internal node, where the former is the variable of the likelihood equation and the coefficients are from the data collected from end-to-end measurement. Our analysis shows such a likelihood equation that actually considers all of the correlations among the descendants in terms of observations regardless of whether the correlations are presented in the dataset collected from an experiment. Because of this, the estimator and alikes are called full likelihood estimators in the literature that has been widely used in loss tomography since most believe using all correlation can lead to accurate estimates. However, this intuitive may not be true if a limited number of measurements are carried out since the usefulness of a correlation depends on whether the data collected from an experiment fits to the model producing the data and whether a correlation considered is completely or partially covered by another [11]. In contrast, the complexity of a full likelihood estimator is directly related to the number of the correlations considered, so is the degree of the polynomial. If the number of descendants is larger than 6, a high degree polynomial is needed to describe the correlations in terms of the observations among the descendants. However, there is no an analytical solution for a fifth or higher degree polynomial equation according to Galois theory. Using an iterative procedure, e.g. the EM, to solve a high degree polynomial certainly affects the scalability of an estimator. Because of this, finding a simple and accurate estimator becomes the key issue of loss tomography that has attracted a consistent attention in the last 10 years. Despite the effort, there has been little progress in this direction. The few explicit estimators, e.g. [8] [12], proposed during this period either perform worse than those using the full likelihood, e.g. [8], or do not provide enough statistical analysis to support themselves, e.g. [12]. This fruitless situation gives rise of two questions: one of them is whether there is an accurate and explicit estimator, while the other is whether the full likelihood estimator is necessary for all data sets. For the former, most believe that accuracy and simplicity cannot coexist whereas there has been little discussion yet for the latter.

To address the questions raised above, we thoroughly analyze the full likelihood estimator presented in [1] and carefully examine the relationship between data collected from an experiment and the model used to describe the data. All of those aim at finding the rational behind of the use of a high degree polynomial as the likelihood equation and look for inspiration that can reduce the degree without loss precision. As previously stated, the rational turns out to be the use of the full likelihood in estimation. With the finding, the correspondences between data and models are identified from the full likelihood estimator. Our focus is then switched to seek an alternative to take advantage of the correspondences. Fortunately, composite likelihood is identified as the replacement of the full likelihood that leads to a number of explicit estimators. The explicit estimators in general perform as good as the full likelihood one in a number of aspect, including variance. In addition, the estimators are not only provide support to real-time applications but also makes it possible to adopt the policy of selecting estimators for data in estimation. The details of the contribution are three-fold:

- the obstacle that blocks the emergence of simple and accurate estimators is identified by analyzing a widely

Weiping Zhu is with University of New South Wales, Australia, email w.zhu@adfa.edu.au

used full likelihood estimator;
- composite likelihood is proposed to replace full likelihood and subsequently a number of explicit estimators are proposed. This enable us to select models for data to have flexibility and accuracy in estimation; and
- the statistical properties of the explicit estimators are presented, including unbiasedness, consistency and uniqueness. More importantly, the asymptotic variance of an explicit estimator is presented that is almost the same as that obtained by a full likelihood estimator.

The rest of the paper is organized as follows. In section 2, we introduce the notations frequently used in the paper and the composite likelihood mentioned above. We then present the analysis of the estimator reported in [1] in section 3. The analysis throws light on the correspondence between data and models in estimation. We then in section 4 divide data and models into a number of groups according to the correlations involved in the observations of the descendants connected to the path being estimated. After that, we pair the data with the models that have the same degree of correlations and obtain a number of composite likelihood equations. Then, a statistic analysis of the proposed estimators is presented in section 5, including the statistic properties of the estimators. The estimators are further evaluated in section 6 by simulations that confirm the contribution listed above. In addition, the robustness of the proposed estimators are discussed in the section that includes the ability to deal with data missing. The last section is devoted to concluding remark.

II. Notation and Composite Likelihood

The symbols that are frequently used in this paper are introduced in this section, where the others will be defined when firstly used.

A. Notation

Let $T = (V, E)$ be a multicast tree used to dispatch probes from a source to a number of receivers. $V = \{v_0, v_1, ...v_m\}$ is a set of nodes and $E = \{e_1, ..., e_m\}$ is a set of directed links that connect the nodes in V to form the multicast tree, where v_0 is the root of the multicast tree. Each node except for v_0 has a parent, let $v_{f(i)}$ be the parent of v_i and e_i is used to connect $v_{f(i)}$ to v_i. If $f_l(i)$ is used to denote the ancestor that is l hops away from node i along the path to the root, we then have $a(i) = \{f(i), f_2(i), \cdot\cdot, f_k(i)\}$ be the ancestors of node i. In addition, we use R to represent all receivers attached to the leaf nodes and use $R(i)$ to represent the receivers attached to $T(i)$ that is a multicast subtree with e_i as its root link. d_i is used to represent the descendants attached to node i that is a nonempty set, where $|d_i|$ is used to denote the number of elements in d_i. Note that a descendant of a node is either a multicast subtree or a link connecting a leaf node. If n probes are sent from v_0 to R, each of them gives rise of an independent realization $X^{(i)}, i = 1, ..., n$, of the pass(loss) process X, $X_k^i = 1, k \in V$ if probe i reaches v_k; otherwise $X_k^i = 0$. The sample $Y = (X_j^{(i)})_{j\in R}^{i\in\{1,..,n\}}$ comprises the data set for estimation that can be divided into sections according to $R(k)$, where Y_k denotes the part of the sample observed by $R(k)$.

If X_k, the pass process of link k, is an independent identical distributed (*i.i.d.*) process and follows a Bernoulli distribution, a set of sufficient statistics have been identified in [13], one for a node that is called the confirmed arrivals at the node and defined as follows:

$$n_k(1) = \sum_{i=1}^{n} \bigvee_{j\in R(k)} y_j^i, \tag{1}$$

where y_j^i denotes the observation of receiver $j, j \in R(k)$, for probe i. $y_j^i = 1$ if the probe is observed by receiver j; otherwise, $y_j^i = 0$.

To understand the principle used by a full likelihood estimator, we need to divide $n_k(1)$ into a number of groups according to the number of descendants observing them. The groups can be divided into two classes: single-observations and co-observations. A single-observation as named refers to the number of probes observed by the receivers attached to a descendant of node k and there are $|d_k|$ single-observations, one for a descendant. In contrast, a co-observation refers to the number of probes simultaneously observed by a number of receivers that are attached to different descendants, called co-observers later. Note that an observer here represents all receivers attached to a descendant, the observer observes a probe if at least one of the receivers observes the probe. The co-observations can be further divided into groups according to the number of co-observers. For node k, there are $2^{|d_k|} - (d_k + 1)$ groups of co-observations. To represent the single-observations and co-observations, a σ-algebra, S, is created over d_k, where $\Sigma_k = S \setminus \emptyset$ is for the single-observations and co-observations. The elements of Σ_k can be divided into a number of groups according to the number of co-observers. For $x, x \in \Sigma_k$, $\#(x)$ is used to denote the number of co-observers in x.

Let γ_j^i be the observation of descendent $T(j)$ for probe i, which is equal to

$$\gamma_j^i = \bigvee_{k\in R(j)} y_k^i$$

Then, we have

$$I_k(x) = \sum_{i=1}^{n} \bigwedge_{j\in x} \gamma_j^i, \ x \in \Sigma_k$$

be the total number of probes observed by the member(s) of x in an experiment. If $\#(x) = 1$, $I_k(x)$ is the single observation of x, otherwise, it is the co-observation of x. Then, we have

$$n_k(1) = \sum_{i=1}^{|d_k|} (-1)^{i-1} \sum_{\substack{\#(x)=i \\ x\in\Sigma_k}} I_k(x), \tag{2}$$

Given (2), we are able to decompose a full likelihood equation, into a number of components.

B. Composite Likelihood

Composite likelihood dates back at least to the pseudo-likelihood proposed by Besag [14] for the applications that have large correlated data sets and highly structured statistical model. Because of the complexity of the applications, it is hard to explicitly describe the correlation embedded in observations. Even if the correlation is describable, the computation complexity often makes inference infeasible in practice [15]. To overcome the difficulty, composite likelihood, instead of full likelihood, is proposed to handle those applications that only consider a number of the correlations within a given data set. The composite likelihood defined in [16] is as follows:

Definition 1:

$$L_c(\theta; y) = \prod_{i \in I} f(y \in C_i; \theta)^{w_i} \tag{3}$$

where $f(y \in C_i; \theta) = f(\{y_j \in Y : y_j \in C_i\}; \theta)$, with $y = (y_1, ..., y_n)$, is a parametric statistical model, $I \subseteq N$, and $\{w_i, i \in I\}$ is a set of suitable weights. The associate composite loglikelihood is $l_c(\theta; y) = \log L_c(\theta, y)$.

As stated, composite likelihoods can be used in parametric estimation. Using the maximum likelihood principle on a composite likelihood function as (3), we have $\triangledown \log L_c(\theta, y)$, the composite likelihood equation. Solving the composite likelihood equation, the maximum composite likelihood estimate is obtained, which under the usual regularity condition is unbiased and the associated maximum composite likelihood estimator is consistent and asymptotically normally distributed [17]. Because of the theoretical properties, composite likelihood has drawn considerable attention in recent years for the applications having complicated correlations, including spatial statistics [18], multivariate survival analysis generalized linear mixed models, and so on [18]. Unfortunately, there has been little work using composite likelihood in network tomography although network tomography is one of the applications that have complex correlations. As far as we know, only [19] proposes an estimator on the basis of pseudo-likelihood for delay tomography and SD traffic matrix tomography.

III. Full Likelihood Estimation

As stated, there has been a lack of discussion about the connection between data and model in a loss rate estimator. To fill the gap, we examine a widely used full likelihood estimator in this section.

A. Full Likelihood and its Components

For the tree topology, the widely used full likelihood estimator is proposed in [1], [2], [3] which is a polynomial as:

$$1 - \frac{\gamma_k}{A_k} = \prod_{j \in d_k} (1 - \frac{\gamma_j}{A_k}). \tag{4}$$

where A_k is the pass rate of the path connecting v_0 to v_k and γ_k is the pass rate of a special subtree that consists of the path from v_0 to v_k and the subtrees rooted at v_k. Despite a number of properties are presented in [1], there is a lack of the necessary and sufficient condition for the correctness of the estimates obtained by (4). Without it, an incorrect estimate can be mistakenly considered a correct one. To remedy this, we present the following theorem that unveils the correspondence between data and models in the likelihood equation.

Theorem 1: The estimate obtained from (4) converges to the true parameter if and only if

1) the true losses occurred on a link is as assumed in [1], i.e. according to the Bernoulli distribution and loss processes of the links are *i.i.d.*; and
2) the observation of $R(k)$ satisfies

$$\forall x, x \in \Sigma_k, I_k(x) \neq 0. \tag{5}$$

Proof: The first condition states such a fact that only if the assumed model is the true model generating the data, the estimates obtained by (4) converges to the true parameter. Based on the conditions, we can write a likelihood function for A_k as

$$L(A_k) = A_k^{n_k(1)} (1 - A_k \beta_k)^{n - n_k(1)} \tag{6}$$

where $1 - \beta_k = \prod_{q \in d_k} (1 - \frac{\gamma_q}{A_k})$ is the loss rate of the subtree rooted at v_k. Turning the above into $\log L(A_k)$, differentiating it and letting the derivative be 0, we have an equation as (4).

Using the empirical probability $\hat{\gamma}_k = \frac{n_k(1)}{n}$ and $\hat{\gamma}_j = \frac{I_k(\{j\})}{n}$ to replace γ_k and γ_j from (4), and then expanding both sides of the equation, we can prove 2) in three steps.

1) If we use (2) to replace $n_k(1)$ from the left hand side (LHS) of (4), the LHS becomes:

$$\begin{aligned} 1 - \frac{\hat{\gamma}_k}{A_k} &= 1 - \frac{n_k(1)}{n \cdot A_k} \\ &= 1 - \frac{1}{n \cdot A_k} [\sum_{i=1}^{|d_k|} (-1)^{i-1} \sum_{\substack{\#(x)=i \\ x \in \Sigma_k}} I_k(x)] \end{aligned} \tag{7}$$

2) If we expand the product term located on the right hand side (RHS) of (4), we have:

$$\prod_{j \in d_k} (1 - \frac{\gamma_j}{A_k}) = 1 - \sum_{i=1}^{|d_k|} (-1)^{i-1} \sum_{\substack{\#(x)=i \\ x \in \Sigma_k}} \frac{\prod_{j \in x} \hat{\gamma}_j}{A_k^i}. \tag{8}$$

3) Deducting 1 from both (7) and (8) and then multiplying the results by A_k, (4) turns to

$$\sum_{i=2}^{|d_k|} (-1)^i \sum_{\substack{\#(x)=i \\ x \in \Sigma_k}} \frac{I_k(x)}{n} = \sum_{i=2}^{|d_k|} (-1)^i \sum_{\substack{\#(x)=i \\ x \in \Sigma_k}} \frac{\prod_{j \in x} \hat{\gamma}_j}{A_k^{i-1}}. \tag{9}$$

It is clear there is a 1-to-1 correspondence between the terms across the equal sign. The correspondence becomes obvious if we rewrite (9) as

$$\sum_{i=2}^{|d_k|} (-1)^i \sum_{\substack{\#(x)=i \\ x \in \Sigma_k}} \Big(\frac{I_k(x)}{n} - \frac{\prod_{j \in x} \hat{\gamma}_j}{A_k^{i-1}} \Big) = 0. \tag{10}$$

Clearly, if one of the $I_k(x) = 0$, (10) could not hold unless the corresponding $\prod_{j\in x}\hat{\gamma}_j = 0$. Then, (4) cannot hold either. The minus sign in the first summation of (10) ensures that each probe is considered once and once only in estimation. ■

B. Insight of Full Likelihood Estimator

Three key points can be drawn from (10) in regards to the full likelihood:

1) the full likelihood relies on both single-observations and co-observations to estimate A_k;
2) the full likelihood requests the presentence of all single- and co-observations although the co-observations are positively correlated; and
3) the full likelihood consists of various likelihoods, each of them represents the correspondence between a group of co-observations and the model generating the co-observations.

The first point indicates the correspondence between single-observations and co-observations, where single-observations and the pass rate, A_k, are used to model the co-observations. This is because although $\prod_{j\in x}\hat{\gamma}_j$ and $\frac{I_k(x)}{n}$ are unbiased estimates for $A_k^{\#(x)}\beta_x$ and $A_k\beta_x$, respectively, where β_x is the product of the pass rates of the subtrees with their root links in x, the former has a smaller variance than the latter [20]. Then, the former is used to model the latter in the likelihood equation. If a single observation is not intersected with others, the single-observation is no longer needed in estimation since there is no corresponded co-observation.

The second point unveils the accuracy of an estimate obtained by the estimator rests on the data set collected from an experiment that requires a complete set of the co-observations defined in Σ_k. Then, a $|d_k|-1$ degree polynomial having $|d_k|$ terms for the degrees from 0 to $|d_k|-1$ becomes inevitable, where each type of the co-observations is considered a dimension and the estimator aims to find a fit within the $|d_k|-1$ dimensions. Instead of dropping some of the correlated observations from estimation, the estimator uses the minus sign of the equation to overcome the possibility of over fitting and ensure a smaller variance for its estimates.

The third point states that each term in the second summation of (10) can be used as an estimator that only considers matching one type of co-observations to the models generating them, i.e. measures the fitness of an A_k in the dimension. Alternatively, we can use a number of the terms in (10) to form an m-dimension space and find an A_k fitting in the space.

IV. Composite Likelihood Estimator

The insights reported in the last section inspire us to search for a different likelihood to construct the likelihood functions. Composite likelihood is then selected as an alternative that allows us to selectively use estimators for data instead of using one for all. To achieve this, a number of theorems and a corollary are presented in this section.

Theorem 2: The maximum composite likelihood equation for A_k that only considers the co-observations of two observers in Y_k is as follows:

$$\Big(\sum_{\substack{\#(x)=2\\ x\in\Sigma_k}}\frac{I_k(x)}{n}\Big)A_k = \sum_{\substack{\#(x)=2\\ x\in\Sigma_k}}\prod_{j\in x}\gamma_j \tag{11}$$

Proof: Let

$$C_i = \{x|\#(x)=2, x\in\Sigma_k\}$$

be the events obtained from Y_k, i.e. the events observed simultaneously by two descendants of v_k. On the basis of (3), we have the pairwise composite likelihood function as follows:

$$L_c(2,A_k;y) = \prod_{i=1}^{|d_k|-1}\prod_{j=i+1}^{|d_k|} f(y_i,y_j;A_k) \tag{12}$$

where $\{i,j\}\in\Sigma_k$ and $w_i = 1$ to have the same weight for each object. The parameter 2 in $L_c(\cdot)$ indicates the likelihood function is for pairwise likelihood. Let $x=\{i,j\}, i,j\in d_k$, we then have a likelihood object as

$$f(y_{R(i)},y_{R(j)};\theta) = A_k^{n_x(1)}(1-A_k\beta_x)^{n-n_x(1)}$$

where $n_x(1) = n_i(1)+n_j(1)-I_k(x)$ and β_x is the combined pass rate of subtrees $T(i)$ and $T(j)$, $i,j\in d_k$, and $1-\beta_x = \prod_{q\in x}(1-\frac{\gamma_q}{A_k})$. Then,

$$L_c(2,A_k;y) = \prod_{\substack{\#(x)=2\\ x\in\Sigma_k}} A_k^{n_x(1)}(1-A_k\beta_x)^{n-n_x(1)} \tag{13}$$

Further, getting the derivative of $\log L_c(2,A_k;y)$, we have

$$\bigtriangledown\log L_c(2,A_k;y) = \sum_{\substack{\#(x)=2\\ x\in\Sigma_k}}\Big[1-\frac{\gamma_k(x)}{A_k}-\prod_{q\in x}(1-\frac{\gamma_q}{A_k})\Big] \tag{14}$$

where $\gamma_k(x)$ denotes the pass rate from v_0 via v_k to the virtual subtree combined by the subtrees inclosed in x. Let (14) be 0, we have the pairwise likelihood equation as (11). ■

It is clear that (11) is equal to set the first term of (10) to 0. Using (13) as the likelihood function, all confirmed arrivals at v_k are taken into account in estimation. The difference between (4) and (11) is the likelihood used in estimation.

If setting each term of (10) to 0, we can have a set of composite likelihood equations, each of them is focused on one type of co-observations. The following corollary expresses the formats of the likelihood functions.

Corollary 1: Let $L_c(1,A_k;y) = 1$, a set of composite likelihood functions are obtained, each considers one type of the $|d_k|-1$ co-observations, from pair-wise to $|d_k|$-wise. The likelihood functions can be expressed in a recursive manner. Let $L_c(i;A_k;y)$ be the $i-wise, 2\le i\le|d_k|$, composite likelihood function that has the following form

$$\begin{aligned} L_c(i;A_k;y) &= \frac{\prod_{\substack{\#(x)=i\\ x\in\Sigma_k}} A_k^{n_x(1)}(1-A_k\beta_x)^{n-n_x(1)}}{\prod_{j=1}^{i-1}L_c(j;A_k;y)} \\ &= \frac{\prod_{\substack{\#(x)=i\\ x\in\Sigma_k}} A_k^{n_x(1)}(1-A_k\beta_x)^{n-n_x(1)}}{\prod_{\substack{\#(y)=i-1\\ y\in\Sigma_k}} A_k^{n_y(1)}(1-A_k\beta_y)^{n-n_y(1)}} \end{aligned} \tag{15}$$

where $n_x(1)$ is the confirmed arrivals at a virtual node x, obtained from Y_x, and β_x is the combined pass rate of the subtrees composed by the members of x.

Proof: It is clear the nominator on the RHS of (15) is the likelihood function considering all co-observations from pairwise to i-wise and the denominator is the likelihood functions from pair-wise to $(i-1)$-wise. The corollary then follows. ∎

There are $|d_k|-1$ composite likelihood functions in (15), each measures the fitness between a type of co-observations and the models generating the co-observations. Since the co-observations considered by each of the composite likelihoods has the same number of co-observers, the likelihood equation obtained is a polynomial that only has two terms, one is an $i-1$ degree term of the estimate and the other is a constant. Then, we have the following theorem.

Theorem 3: Each of the composite likelihood equations obtained from (15) is an explicit estimator of A_k. Let $A_k(i)$ be the explicit estimator for *i-wise* co-observations that returns an estimate of A_k. We then have

$$A_k(i) = \Big(\frac{\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \prod_{j\in x}\gamma_j}{\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \frac{I_k(x)}{n}}\Big)^{\frac{1}{i-1}}, i\in\{2,..,|d_k|\}. \quad (16)$$

Proof: Firstly, writing (15) into a log-likelihood, we then differentiate the log-likelihood and let the derivative be 0. Finally, the likelihood equation for the i-wise likelihood function is presented as (16). ∎

In the rest of the paper, we use $A_k(i)$ to refer the $i-wise$ estimator and $\hat{A}_k(i)$ to refer the estimate obtained by $A_k(i)$, where the i is called the index of the estimator.

A. Weighted Mean

Let z_j be the pass rate of descendant j. If knowing z_j, (16) can be written as

$$A_k(i)^{i-1} = \sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \frac{\prod_{j\in x} z_j}{\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} (\prod_{j\in x} z_j)} A_x^{i-1}. \quad (17)$$

where A_x^{i-1} is equal to $\frac{\prod_{j\in x}\gamma_j}{\prod_{j\in x} z_j}$. Then, $\hat{A}_k(i)^{i-1}$ is a weighted mean of $A_x^{i-1}, x\in\Sigma_k \bigwedge \#(x)=i$, and $\hat{A}_k(i)$ is a weighted power mean. The weight assigned to A_x^{i-1} depends on the pass rates of the subtrees in x. The A_x that has a high weight contributes more to the mean than those having low weights. Since z_j is not observable directly, γ_j and $I_x(x)$ are used in (16) to achieve the same effect as the weighted power mean. Because of this, the estimators share a similar statistical properties as the weighted power mean.

V. Properties of Composite Likelihood Estimator

This section is devoted to evaluate the statistical properties of the $|d_k|-1$ estimators presented in (16), include unbiasedness, consistency, uniqueness, and the robustness. The following theorems or lemmas are for the properties. Firstly, we have

Lemma 1: Let

$$lm_k(x) = \frac{\prod_{j\in x}\gamma_j}{\frac{I_k(x)}{n}} \qquad x\in\Sigma_k, \#(x)\geq 2,$$

$lm_k(x)$ is an unbiased estimate of $A_k^{\#(x)-1}$.

Proof: Let $\hat{n}_k(1)$ be the number of probes reaching v_k and let $\overline{z_j} = \frac{n_j(1)}{\hat{n}_k(1)}$ be the sample mean of the pass rate of descendant j.

$$\begin{aligned}
E(lm_k(x)) &= E\Big(\frac{\prod_{j\in x}\gamma_j}{\frac{I_k(x)}{n}}\Big) \\
&= E\Big(\frac{\prod_{j\in x}\frac{n_j(1)}{n}}{\frac{\sum_{i=1}^{n}\bigwedge_{j\in x} y_j^i}{n}}\Big) \\
&= E\Big(\frac{(\frac{\hat{n}_k(1)}{n})^{\#(x)}\prod_{j\in x}\frac{n_j(1)}{\hat{n}_k(1)}}{\frac{\hat{n}_k(1)}{n}\frac{\sum_{i=1}^{\hat{n}_k(1)}\prod_{j\in x} z_j^i}{\hat{n}_k(1)}}\Big) \\
&= E\Big((\frac{\hat{n}_k(1)}{n})^{\#(x)-1}\Big)E\Big(\frac{\prod_{j\in x}\overline{z_j}}{\prod_{j\in x} z_j}\Big) \\
&= E\Big((\frac{\hat{n}_k(1)}{n})^{\#(x)-1}\Big)E\Big(\prod_{j\in x}\frac{\overline{z_j}}{z_j}\Big) \\
&= E\Big(\bar{A_k}^{\#(x)-1}\Big) \\
&= A_k^{\#(x)-1} \qquad (18)
\end{aligned}$$

∎

Lemma 1 shows that although $lm_k(x)$ only considers a part of the observations obtained from an experiment, it can be an unbiased estimator of $A_k^{\#(x)-1}$. To simplify the following discussion, we call $lm_k(\cdot)$ local estimator. Accordingly, the estimate obtained by $lm_k(\cdot)$ is called local estimate and denoted by $\widehat{lm}_k(\cdot)$. Based on lemma 1 we have

Theorem 4: $(\widehat{lm}_k(x))^{\frac{1}{\#(x)-1}}, x\in\Sigma_k$ is an unbiased estimate of A_k.

Proof: This can be obtained directly for lemma 1. ∎

Since $A_k(|d_k|)$ is the $|d_k|-1$-th root of $lm_k(d_k)$, we can conclude that $A_k(|d_k|)$ is an unbiased estimator. Apart from $A_k(|d_k|)$, we are not able to prove other $A_k(i)$s are unbiased estimators since we are not able to prove

$$E\Big(\Big(\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \frac{\prod_{j\in x}\overline{z_j}}{\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}}\prod_{j\in x} z_j}\Big)^{\frac{1}{i-1}}\Big) = 1$$

if $n<\infty$. Nevertheless, we have

Theorem 5: The estimators presented in (16) are asymptotic unbiased estimators.

Proof: According to lemma 1,

$$\hat{A}_k(i) = \bar{A_k}\Big(\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \frac{\prod_{j\in x}\overline{z_j}}{\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}}\prod_{j\in x} z_j}\Big)^{\frac{1}{i-1}} \quad (19)$$

Then,

$$\begin{aligned}
&\lim_{n\to\infty} E(\hat{A}_k(i)) = \\
&\quad \lim_{n\to\infty} E(\bar{A}_k) E\Big(\Big(\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \frac{\prod_{j\in x} \overline{z_j}}{\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \prod_{j\in x} z_j}\Big)^{\frac{1}{i-1}}\Big) \\
&\quad = A_k E\Big(\lim_{n\to\infty} \Big(\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \frac{\prod_{j\in x} \overline{z_j}}{\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \prod_{j\in x} z_j}\Big)^{\frac{1}{i-1}}\Big) \\
&\quad = A_k E\Big(\Big(\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \frac{\prod_{j\in x} z_j}{\sum_{\substack{x\in\Sigma_k \\ \#(x)=i}} \prod_{j\in x} z_j}\Big)^{\frac{1}{i-1}}\Big) \\
&\quad = A_k
\end{aligned} \tag{20}$$

The theorem follows. ∎

Further, we have the following lemma before proving $\hat{A}_k(i)$ is a consistent estimate.

Lemma 2: 1) $\widehat{lm}_k(x)$ is a consistent estimate of $A_k^{\#(x)-1}$; and

2) $\widehat{lm}_k(x)^{\frac{1}{\#(x)-1}}$ is a consistent estimate of A_k.

Proof:

1) Lemma 1 shows that $\widehat{lm}_k(x)$ is equivalent to the first moment of $A_k^{\#(x)-1}$. Then, according to the law of large number, $\widehat{lm}_k(x) \to A_k^{\#(x)-1}$.
2) From the above and the continuity of $lm_k(x)$ on the values of $\gamma_j, j \in x$ and $I_k(x)/n$ generated as A_k ranges over its support set, the result follows.

∎

Then, we have

Theorem 6: $\hat{A}_k(i)$ is a consistent estimate of A_k.

Proof: As stated, $A_k(i)$ is an estimator using the weighted power mean in estimation. Let $S_k(i) = \{x | x \in \Sigma_k \bigwedge \#(x) = i\}$, we have

$$\min_{x\in S_k(i)} \widehat{lm}_k(x)^{\frac{1}{i-1}} \le \hat{A}_k(i) \le \max_{x\in S_k(i)} \widehat{lm}_k(x)^{\frac{1}{i-1}}. \tag{21}$$

Then, from lemma 2, the theorem follows. ∎

Further, we have the following for uniqueness.

Theorem 7: If

$$\sum_{\substack{\#(x)=i \\ x\in\Sigma_k}} \prod_{j\in x} \gamma_j < \sum_{\substack{\#(x)=i \\ x\in\Sigma_k}} \frac{I_k(x)}{n},$$

there is only one solution in $(0,1)$ from $A_k(i), 2 \le i \le |d_k|$.

Proof: This is obvious. ∎

As [1], [8] using a linear function to model the loss rate of a link, the Delta method can be used to prove that the asymptotic variance of $A_k(i)$ is the same as that of the full likelihood estimate, to the first order of the maximum loss rate of the link in E. Firstly, a number of symbols introduced in [1] and [8] for theorem 5 and lemma 1, respectively, are presented, which will be used in the proof of a lemma and a theorem later in this section, i.e. lemma 3 and theorem 8. Among the symbols, $\alpha_k, k \in E$ is for the pass rate of link k, $\bar{\alpha}_k = 1-\alpha_k$ is for the loss rate of link k, and $\|\bar{\alpha}\| = \max_{k\in E} |\bar{\alpha}_k|$. In addition, we have γ_x for the pass rate of a special tree consisting of the path from node 0 to v_k and the subtrees rooted at node k and with their root links in x. In addition, let $t_x = \sum_{j\in x} \bar{\alpha}_j, x \in \Sigma_k$. Then, we have the following lemma that generalizes Lemma 1 in [8]:

Lemma 3: 1) $Cov(\gamma_x, \gamma_y) = \gamma_x(1-\gamma_y)$ if $y \subseteq x$ and $x, y \in \Sigma_k$; otherwise $\gamma_{x\vee y} - \gamma_x\gamma_y$ if $x, y \in \Sigma_k$ and $y \nsubseteq x$ or $x \nsubseteq y$.

2) $Cov(\gamma_x, \gamma_y) = s_k + \bar{\alpha}_y + O(\|\bar{\alpha}\|^2)$ if $y \subseteq x$ and $x, y \in \Sigma_k$, $Cov(\gamma_x, \gamma_y) = s_k + t_{x\wedge y} + O(\|\bar{\alpha}\|^2)$ if $x, y \in \Sigma_k$ and $y \nsubseteq x$ or $x \nsubseteq y$.

Proof: x and y here can be considered virtual nodes if $\#(x)$ or $\#(y)$ is larger than 1, where $\hat{\gamma}_x = \frac{I_k(x)}{n}, x \in \Sigma_k$ and $\hat{\gamma}_y = \frac{I_k(y)}{n}, y \in \Sigma_k$. Since the loss process is a 0-1 distribution, we have

$$\begin{aligned}
E(\hat{\gamma}_x\hat{\gamma}_y) &= \Big[\prod_{j\in x\vee y} P(Y_j = 1 | X_k = 1)\Big] P(X_k = 1) \\
&= \Big[\prod_{j\in x\vee y} \frac{P(Y_j = 1 \bigwedge X_k = 1)}{P(X_k = 1)}\Big] P(X_k = 1) \\
&= \frac{\prod_{j\in x\vee y} \gamma_j}{A_k^{\#(x\vee y)-1}}
\end{aligned} \tag{22}$$

if $x \not\supseteq y$ and $y \not\supseteq x$. Then,

$$\begin{aligned}
Cov(\gamma_x\gamma_y) &= E(\gamma_x\gamma_y) - E(\gamma_x)E(\gamma_y) \\
&= \frac{\prod_{j\in x\vee y} \gamma_j}{A_k^{\#(x\vee y)-1}} - \frac{\prod_{j\in x} \gamma_j}{A_k^{\#(x)-1}} \cdot \frac{\prod_{i\in y} \gamma_i}{A_k^{\#(y)-1}} \\
&= \frac{\prod_{j\in x} \gamma_j \prod_{i\in y} \gamma_i}{A_k^{\#(x)+\#(y)-2}} \Big(\frac{1}{A_k^{\#(x\wedge y)+1}} - 1\Big) \\
&= \gamma_x\gamma_y \Big(\frac{1}{A_k^{\#(x\wedge y)+1}} - 1\Big) \\
&= \gamma_{x\vee y} - \gamma_x\gamma_y
\end{aligned} \tag{23}$$

If $x, y \in \Sigma_k$ and $y \subset x$, we have

$$E(\hat{\gamma}_x\hat{\gamma}_y) = E(\hat{\gamma}_x) = \hat{\gamma}_x \tag{24}$$

and then

$$Cov(\gamma_x, \gamma_y) = \gamma_x(1-\gamma_y) \tag{25}$$

2) According to Lemma 1 in [8], where $A_k = 1 - s_k + O(\| \bar{\alpha} \|^2)$ and $\gamma_k = 1 - s_k + O(\| \bar{\alpha} \|^2)$, $\gamma_x = A_k \prod_{j\in x} \frac{\gamma_j}{A_k}$, $\gamma_x = 1 - s_k - t_x + O(\|\bar{\alpha}\|^2)$. Using the values in (23), we have

$$Cov(\gamma_x, \gamma_y) = s_k + t_{x\wedge y} + O(\|\bar{\alpha}\|)^2 \tag{26}$$

for $x, y \in \Sigma_k$ and $y \nsubseteq x$ or $x \nsubseteq y$, where $s_k = \sum_{j\in a(k)} \bar{\alpha}_j$. For $x, y \in \Sigma_k$ and $y \subset x \wedge \#(y) = 1$. Then,

$$\begin{aligned}
&Cov(\gamma_x, \gamma_y) = \gamma_x(1-\gamma_y) \\
&= (1 - s_k - t_x + O(\|\bar{\alpha}\|)^2)(1 - (1 - s_y + O(\|\bar{\alpha}\|)^2)) \\
&= (1 - s_k - t_x + O(\|\bar{\alpha}\|)^2)(s_k + \bar{\alpha}_y + O(\|\bar{\alpha}\|)^2) \\
&\doteq s_k + \bar{\alpha}_y + O(\|\bar{\alpha}\|)^2.
\end{aligned} \tag{27}$$

∎

Given lemma 3, we have the following theorem for the asymptotic variance of $\hat{A}_k(i)$.

Theorem 8: As $n \rightarrow \infty$, $n^{\frac{1}{2}}(\hat{A}_k(i) - A_k)$ has an asymptotically Gaussian distribution of mean 0 and variance $v_k^E(A_k(i))$, where $v_k^E(A_k(i)) = s_k + O(\| \bar{\alpha} \|^2)$.

Proof: Using the central limit theorem, we can prove $n^{\frac{1}{2}}(A_k(i) - A_k)$ is an asymptotically Gaussian distribution of mean 0 and variance v_k^E as $n \rightarrow \infty$. We here focus on prove the variance. Let

$$D(i) = \{x | \#(x) = i, x \in \Sigma_k\}.$$

that has

$$|D(i)| = \binom{|d_k|}{i}$$

elements. By the Delta method, $n^{\frac{1}{2}}(\hat{A}_k(i) - A_k)$ converges in distribution to a zero mean Gaussian random variable with variance

$$v_k^E(A_k(i)) = \bigtriangledown A_k(i) \cdot C^E \cdot \bigtriangledown A_k(i)^T$$

where $\bigtriangledown A_k(i)$ is a derivative vector obtained as follows:

$$\bigtriangledown A_k(i) = \Big(\Big\{ \frac{\partial A_k(i)}{\partial \gamma_x} : x \in D(i) \Big\}, \Big\{ \frac{\partial A_k(i)}{\partial \gamma_j} : j \in D(1) \Big\} \Big). \quad (28)$$

C^E is a square covariance matrix in the form of

$$C^E = \begin{pmatrix} \{Cov(\gamma_x, \gamma_y)\} & \{Cov(\gamma_x, \gamma_j)\} \\ \{Cov(\gamma_x, \gamma_j)\} & \{Cov(\gamma_j, \gamma_{j'})\} \end{pmatrix} \quad (29)$$

Each of the four components is a matrix, where $x, y \in D(i)$ and $j, j' \in D(1)$. To calculate $\bigtriangledown A_k(i)$ and derive the first order of v_k^E in $\bar{\alpha}$, let $de(i)$ denote the denominator of $A_k(i)$ and $no(i)$ denote the nominator. In addition, let $no(i,j) \subset no(i), j \in d_k$ be the terms of $no(i)$ that have γ_j. Then, we have

$$\bigtriangledown A_k(i) = \frac{A_k(i)}{i-1} \Big(\{ \frac{-1}{de(i)} : j \in D(i) \}, \{ \frac{no(i,j)}{\gamma_j \cdot no(i)} : j \in D(1) \} \Big)$$

The number of terms in the denominator and the nominator of $A_k(i)$ is the same, i.e. $|de(i)| = |no(i)|$ and $|no(i,j)| = \binom{|d_k|-1}{i-1}$ and $var(\gamma_x) = \gamma_x(1-\gamma_x) = s_k + t_x + O(\|\bar{\alpha}\|^2)$. Inserting the values and those obtained in Lemma 3 into $\bigtriangledown A_k(i)$ and C^E, we have

$$\bigtriangledown A_k(i) = \frac{1}{i-1} \Big(\{ \frac{-1}{|D(i)|} : j \in D(i) \}, \{ \frac{\binom{|d_k|-1}{i-1}}{|D(i)|} : j \in D(1) \} \Big) + O(\| \bar{\alpha} \|)$$

and

$$C^E = s_k + \begin{pmatrix} \{t_{x \wedge y} : x, y \in D(i)\} & \{\bar{\alpha}_j : j \in D(1)\} \\ \{\bar{\alpha}_j, j \in D(1)\} & Diag\{\bar{\alpha}_j : j \in D(1)\} \end{pmatrix} + O(\|\bar{\alpha}\|^2) \quad (30)$$

$\{t_{x \wedge y} : x, y \in D(i)\}$ is a symmetric matrix, where the diagonal entries are equal to $t_x, x \in D(i)$ since $t_{x \wedge x} = t_x$. The sum of each column in $\{t_{x \wedge y} : x, y \in D(i)\}$ is equal to

$$\sum_{j \in x} \binom{|d_k|-1}{i-1} \bar{\alpha}_j.$$

The bottom left matrix has $|D(1)|$ columns, there are $\#(x)$ entries in a column that equal to $\bar{\alpha}_j, j \in x$, respectively. The top right matrix is the transpose of the bottom left one that has $\binom{|d_k|-1}{i-1}$ entries in a column that all equal to $\bar{\alpha}_j$; the bottom right matrix is a diagonal matrix that has its diagonal entries equaling to $\bar{\alpha}_j, j \in D(1)$. Let M denote the the matrix in (30). We have $\bigtriangledown A_k(i) \cdot M \cdot \bigtriangledown A_k(i)^T = 0$ Then, the theorem follows. ■

Compared with theorem 3 (i) in [8], theorem 8 shows that the asymptotic variance of $A_k(i)$ is the same as (4), to the first order in the pass rate of a link.

Corollary 2: Under the same condition as that in theorem 8, the asymptotic variance of $lm_k(x)$ is equal to $s_k + O(\| \bar{\alpha} \|^2)$.

Proof: $lm_k(x)$ can be considered a special $A_k(i)$ that has a two-element $\bigtriangledown lm_k(x)$ and a four-element C^E. Using the same procedure as that in the proof of theorem 8, we have the corollary. ■

A. Example

we use an example to illustrate that composite estimators has an asymptotic variance that is very close to that obtained by (4). The setting used by the example is identical to the one presented in [8], where v_k has three descendants with a pass rate of α and the pass rate from v_0 to v_k is also equal to α. Three estimators are considered, which are $A_k(2)$, $A_k(3)$ and (4). Note that

$$A_k(3) = \Big(\frac{\gamma_1 \gamma_2 \gamma_3}{\frac{I_k(\{1,2,3\})}{n}} \Big)^{1/2}$$

is the same as the explicit estimator proposed in [8]. On the other hand,

$$A_k(2) = \frac{\gamma_1\gamma_2 + \gamma_1\gamma_3 + \gamma_2\gamma_3}{\frac{I_k(\{1,2\})}{n} + \frac{I_k(\{1,3\})}{n} + \frac{I_k(\{2,3\})}{n}} \quad (31)$$

where

$$v_k^E(A_k(2)) = \nabla A_k(2) C^E \nabla A_k(2) \quad (32)$$

and

$$D(2) = \{x | \#(x) = 2, x \in \Sigma_k\}.$$

We then have C^E that is the $2 * |d_k|$ dimensional square covariance matrix in the form of (29).

Let $\gamma_{\{1,2\}} = \frac{I_k(\{1,2\})}{n}$, $\gamma_{\{2,3\}} = \frac{I_k(\{2,3\})}{n}$, and $\gamma_{\{1,3\}} = \frac{I_k(\{1,3\})}{n}$. Based on the above setting we have

$\gamma_{\{i,j\}} = \alpha^3, i, j \in \{1,2,3\}$ and $\gamma_l = \alpha^2, l \in \{1,2,3\}$. Then, the top-left of C^E is

$$\begin{pmatrix} \alpha^3(1-\alpha^3) & \alpha^6(\frac{1}{\alpha^2}-1) & \alpha^6(\frac{1}{\alpha^2}-1) \\ \alpha^6(\frac{1}{\alpha^2}-1) & \alpha^3(1-\alpha^3) & \alpha^6(\frac{1}{\alpha^2}-1) \\ \alpha^6(\frac{1}{\alpha^2}-1) & \alpha^6(\frac{1}{\alpha^2}-1) & \alpha^3(1-\alpha^3) \end{pmatrix}$$

the top-right is

$$\begin{pmatrix} \alpha^3(1-\alpha^2) & \alpha^3(1-\alpha^2) & \alpha^6(\frac{1}{\alpha^2}-\frac{1}{\alpha}) \\ \alpha^3(1-\alpha^2) & \alpha^6(\frac{1}{\alpha^2}-\frac{1}{\alpha}) & \alpha^3(1-\alpha^2) \\ \alpha^6(\frac{1}{\alpha^2}-\frac{1}{\alpha}) & \alpha^3(1-\alpha^2) & \alpha^3(1-\alpha^2) \end{pmatrix}$$

the bottom-left is

$$\begin{pmatrix} \alpha^3(1-\alpha^2) & \alpha^3(1-\alpha^2) & \alpha^5(\frac{1}{\alpha}-1) \\ \alpha^3(1-\alpha^2) & \alpha^5(\frac{1}{\alpha}-1) & \alpha^3(1-\alpha^2) \\ \alpha^5(\frac{1}{\alpha}-1) & \alpha^3(1-\alpha^2) & \alpha^3(1-\alpha^2) \end{pmatrix}$$

and the bottom-right is

$$\begin{pmatrix} \alpha^2(1-\alpha^2) & \alpha^4(\frac{1}{\alpha}-1) & \alpha^4(\frac{1}{\alpha}-1) \\ \alpha^4(\frac{1}{\alpha}-1) & \alpha^2(1-\alpha^2) & \alpha^4(\frac{1}{\alpha}-1) \\ \alpha^4(\frac{1}{\alpha}-1) & \alpha^4(\frac{1}{\alpha}-1) & \alpha^2(1-\alpha^2) \end{pmatrix}$$

In addition, we have

$$\nabla A_k(2) = (\frac{-1}{3\alpha^2}, \frac{-1}{3\alpha^2}, \frac{-1}{3\alpha^2}, \frac{2}{3\alpha}, \frac{2}{3\alpha}, \frac{2}{3\alpha}).$$

Using (32), we have

$$v_k^E(A_k(2)) = \bar{\alpha} - \frac{2}{3}\bar{\alpha}^2 + O(\|\bar{\alpha}\|^3).$$

Comparing this with the results presented in [8] for $A_k(3)$ and $\hat{A}_k$, where

$$v_k^E(A_k(3)) = \bar{\alpha} - \frac{\bar{\alpha}^2}{4} + O(\|\bar{\alpha}\|^3)$$

and

$$v_k^E(\hat{A}_k) = \bar{\alpha} - \bar{\alpha}^2 + O(\|\bar{\alpha}\|^3),$$

$A_k(2)$ performs better than $A_k(3)$ in terms of asymptotic variance. This shows that an estimator consisting of more terms may perform better than the others with less terms. Note that the full likelihood estimator is the minimum variance unbiased estimator that has been proved in [13].

VI. Estimator Evaluation

According to (19), the quality of $A_k(i)$ depends on

$$f(i) = \Big(\sum_{\substack{x \in \Sigma_k \\ \#(x)=i}} \frac{\prod_{j \in x} \overline{z_j}}{\sum_{\substack{x \in \Sigma_k \\ \#(x)=i}} \prod_{j \in x} z_j} \Big)^{\frac{1}{i-1}}, \qquad (33)$$

which measures the fitness between $i-wise$ co-observations and $i-wise$ model although $f(i)$ is not observable. If $f(i)$ equals to 1, the data fits to the model perfectly. As stated, each type of co-observations can be considered a dimension, the estimators presented in (16) measure the fitness in different dimensions. Although the fitness in one dimension could not ensure the fitness in the others, theorem 6 ensures that $\forall i, i \in \{2,..,|d_k|\}, f(i) \to 1$ as $n \to \infty$. Thus, as $n \to \infty$, all $\hat{A}_k(i)$s coverage to the true value and there is litter difference between them.

If $n < \infty$, the estimates obtained by some of the estimators in (16) may be better than that obtained by the full likelihood estimator. We can use the composite Kullback-Leigh divergence to find the best estimator for a data set, where Akaike information criterion (AIC) [21] is computed for each of the estimators.

To confirm the above, three rounds of simulations are conducted in various setting, where five estimators are evaluated, i.e. the full likelihood, pairwise likelihood, i.e. $A_k(2)$), triple-wise likelihood, i.e. $A_k(3)$, a pair local, i.e $lm_k(2)$, and a triple local, i.e. $lm_k(3)$ are compared against each other and the results are presented in a number of tables. The number of samples used in the simulations starts from 300 and end at 9900 in a step of 300. For each sample size, 20 experiments are carried out to obtain the means and variance. In the tables, we only present a part of the results in the tables, where all of the means and variance for the samples varying from 300 to 3000 are included while two samples, i.e. 4800 and 9900, are presented for the other samples that are larger than 3000. Table I is the results obtained from a tree with 8 descendants, where the loss rate of a link is set to 1%. In general, when the sample is small, the estimates obtained by all estimators are drifted away from the true value that indicates the data obtained is not enough. Once the sample size reaches 2000, the estimates approach to the true value that shows there is enough information to make an accurate estimate. All of the estimators achieve the same outcome with the increase of samples. However, with the increase of samples, the variance reduces slowly although there are a number of exceptions. The results also show that if the loss rate is lower, such as 1%, there is little difference between the estimators in terms of the means of the estimated obtained and the variance of the estimates.

To investigate the impact of different loss rates on estimation, another round experiment is carried out on the same network topology, where 6 of the descendant links have their loss rates equal to 1% and the other two have 5%. The paired local estimator uses the co-observation that consists of one from each class to estimate the loss rate, while the triple local estimator uses the co-observation consisting of two of 1% links and one of 5%. The results are presented in Table II. Compared with Table I, there is little difference among the full likelihood, $A_k(2)$ and $A_k(3)$ in terms of the means and variances obtained by them. In contrast, there is a slight difference between the two round simulations for the two local estimators, where the variance of the second round in most cases is slightly larger than that of the first round. This is because of the higher loss rates on some of the descendant links that certainly has its impact on the variance of the estimate of the root link, in particular if the sample size is small. Comparing the two local estimators, one is able

Estimators	Full Likelihood		Pair Likelihood		Triple Likelihood		Single Pair		Single Triple	
samples	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
300	0.0088	1.59E-05	0.0088	1.59E-05	0.0088	1.64E-05	0.0087	1.59E-05	0.0087	1.61E-05
600	0.0089	1.12E-05	0.0089	1.12E-05	0.0089	1.13E-05	0.0089	1.10E-05	0.0088	1.12E-05
900	0.0092	7.76E-06	0.0092	7.82E-06	0.0091	7.84E-06	0.0092	7.90E-06	0.0092	8.15E-06
1200	0.0095	6.13E-06	0.0095	6.13E-06	0.0094	6.17E-06	0.0095	6.16E-06	0.0095	5.97E-06
1500	0.0096	4.55E-06	0.0096	4.55E-06	0.0096	4.80E-06	0.0096	4.78E-06	0.0096	4.33E-06
1800	0.0096	1.82E-06	0.0096	1.81E-06	0.0096	1.92E-06	0.0097	1.92E-06	0.0096	1.90E-06
2100	0.0097	3.14E-06	0.0097	3.11E-06	0.0097	3.14E-06	0.0097	3.02E-06	0.0097	3.08E-06
2400	0.0100	1.32E-06	0.0100	1.32E-06	0.0100	1.36E-06	0.0100	1.29E-06	0.0099	1.28E-06
2700	0.0100	1.72E-06	0.0100	1.72E-06	0.0100	1.74E-06	0.0100	1.81E-06	0.0100	1.83E-06
3000	0.0102	2.96E-06	0.0102	2.97E-06	0.0102	3.01E-06	0.0102	3.04E-06	0.0102	2.95E-06
4800	0.0103	1.74E-06	0.0103	1.74E-06	0.0103	1.74E-06	0.0103	1.75E-06	0.0103	1.81E-06
9900	0.0099	8.18E-07	0.0099	8.23E-07	0.0099	8.20E-07	0.0099	8.05E-07	0.0099	8.60E-07

TABLE I
SIMULATION RESULT OF A 8-DESCENDANT TREE WITH LOSS RATE=1%

Estimators	Full Likelihood		Pair Likelihood		Triple Likelihood		Single Pair		Single Triple	
samples	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
300	0.0088	1.59E-05	0.0089	1.64E-05	0.0089	1.68E-05	0.0091	2.36E-05	0.0088	1.95E-05
600	0.0089	1.12E-05	0.0089	1.14E-05	0.0089	1.16E-05	0.0088	1.46E-05	0.0089	1.26E-05
900	0.0091	7.76E-06	0.0091	7.80E-06	0.0091	7.83E-06	0.0092	9.74E-06	0.0091	8.67E-06
1200	0.0094	6.13E-06	0.0094	6.16E-06	0.0094	6.18E-06	0.0096	7.09E-06	0.0095	6.16E-06
1500	0.0096	4.55E-06	0.0096	4.72E-06	0.0096	4.81E-06	0.0097	4.36E-06	0.0096	4.45E-06
1800	0.0096	1.82E-06	0.0096	1.90E-06	0.0096	1.95E-06	0.0096	2.45E-06	0.0096	1.97E-06
2100	0.0097	3.14E-06	0.0097	3.11E-06	0.0097	3.11E-06	0.0098	3.39E-06	0.0097	3.04E-06
2400	0.0099	1.32E-06	0.0100	1.34E-06	0.0100	1.35E-06	0.0101	1.64E-06	0.0100	1.44E-06
2700	0.0100	1.72E-06	0.0100	1.69E-06	0.0100	1.67E-06	0.0101	2.11E-06	0.0100	1.90E-06
3000	0.0102	2.96E-06	0.0102	2.93E-06	0.0102	2.91E-06	0.0103	2.83E-06	0.0102	2.87E-06
4800	0.0103	1.74E-06	0.0104	1.74E-06	0.0104	1.74E-06	0.0104	2.06E-06	0.0104	2.01E-06
9900	0.0099	8.18E-07	0.0099	8.30E-07	0.0099	8.36E-07	0.0099	9.78E-07	0.0099	9.11E-07

TABLE II
SIMULATION RESULT OF A 8-DESCENDANT TREE, 6 OF THE 8 HAVE LOSS RATE=1% AND THE OTHER 2 HAVE LOSS RATE=5%

Estimators	Full Likelihood		Pair Likelihood		Triple Likelihood		Single Pair		Single Triple	
samples	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
300	0.0503	2.15E-04	0.0504	2.15E-04	0.0505	2.14E-04	0.0508	2.18E-04	0.0505	2.16E-04
600	0.0503	8.23E-05	0.0503	8.21E-05	0.0503	8.19E-05	0.0504	8.24E-05	0.0503	8.27E-05
900	0.0511	5.85E-05	0.0511	5.81E-05	0.0511	5.79E-05	0.0512	5.79E-05	0.0512	5.88E-05
1200	0.0506	4.93E-05	0.0506	4.97E-05	0.0507	4.99E-05	0.0507	4.85E-05	0.0507	4.93E-05
1500	0.0502	2.24E-05	0.0502	2.24E-05	0.0502	2.23E-05	0.0503	2.33E-05	0.0502	2.32E-05
1800	0.0500	3.89E-05	0.0500	3.85E-05	0.0500	3.83E-05	0.0501	3.91E-05	0.0500	3.94E-05
2100	0.0507	1.16E-05	0.0507	1.19E-05	0.0507	1.20E-05	0.0507	1.09E-05	0.0507	1.13E-05
2400	0.0510	1.40E-05	0.0510	1.43E-05	0.0510	1.44E-05	0.0510	1.40E-05	0.0510	1.43E-05
2700	0.0507	1.31E-05	0.0507	1.34E-05	0.0507	1.35E-05	0.0508	1.35E-05	0.0507	1.34E-05
3000	0.0508	6.65E-06	0.0508	6.98E-06	0.0508	7.14E-06	0.0508	6.79E-06	0.0508	6.85E-06
4800	0.0498	1.09E-05	0.0498	1.10E-05	0.0498	1.10E-05	0.0498	1.11E-05	0.0498	1.11E-05
9900	0.0496	5.35E-06	0.0496	5.38E-06	0.0497	5.40E-06	0.0496	5.48E-06	0.0496	5.48E-06

TABLE III
SIMULATION RESULT OF A 8-DESCENDANT TREE, THE LOSS RATE OF THE ROOT LINK=5%, 4 OF THE 8 HAVE LOSS RATE=1% AND THE OTHER 4 HAVE LOSS RATE=5%

to notice that the local estimator considering the triple co-observation performs a slightly better than that considering the paired co-observation in terms of variance. This indicates that a local estimator is sensitive to the co-observation selected for its estimation and the co-observation involving more members can overcome the turbulence created by short term losses in some degree and provide accurate estimates.

To further investigate the impact of loss rates on estimation, we conduct another round simulation, where we increase the loss rate of the root link from 1% to 5%, set the loss rates of four descendent links to 5% and the rest to 1%. The result is presented in Table III, where the local estimators consider the observations obtained from the descendants that have 1% loss rate. The obvious difference between the result and the previous two is the variance that is a magnitude higher. This indicates a large variance is expected for the estimates of a long path that traverses a number of serially connected links since the loss rate of a path is proportional to the number of links in the path. To reduce the variance, we need to send more probes.

A. Classification

The simulation study shows the stability of the explicit estimators proposed in this paper, where $A_k(2)$ and $A_k(3)$ perform as good as the full likelihood estimator proposed in [1] in all of the settings. Considering the differences in terms of the likelihood used by the estimators proposed so far, we can have a classification for them, where the full likelihood estimators and the local likelihood ones stand at the two extremes. The full likelihood estimator indiscriminately

considers all of the co-observations in estimation and uses the average of the co-observations to overcome the possible surge created by a few individual co-observations. Because of this, we consider it a blind estimator. In contrast, a local estimator is focused on a specific co-observation in its estimation, the accuracy of an estimate then rests on the information provided by the co-observation, which can be venerable if the sample size is small. We call them the specific estimators. The composite likelihood estimators proposed in this paper stay in between the two extremes that can consider a type of co-observations as 16 or a few types of the co-observations, or a part of the co-observations. If there is no special bias to a type of co-observations, the composite likelihood estimators performs similar to the full likelihood one since there is enough redundance in the information provided by co-observations. .

B. Other Estimators

Apart from the estimators in (16), there are still a large number of composite likelihood estimators for loss tomography. For instance, we proposed a number of estimators in [12] that divide the descendants of a node into groups and consider a group as a virtual descendant of the node. If d_k is divided into two groups, we can divide Y_k into two groups accordingly and let k_1 and k_2 denote the two virtual descendants. Then, as (13) we can have $n_{k1}(1)$ and $n_{k2}(1)$ as the confirmed arrivals at the two virtual nodes k_1 and k_2. Then, we have a composite likelihood estimator as

$$1 - \frac{\hat{\gamma}_k}{A_k} = (1 - \frac{\hat{\gamma}_{k1}}{A_k})(1 - \frac{\hat{\gamma}_{k2}}{A_k}) \tag{34}$$

where $\hat{\gamma}_k = \frac{n_k(1)}{n}$, $\hat{\gamma}_{k1} = \frac{n_{k1}(1)}{n}$, and $\hat{\gamma}_{k2} = \frac{n_{k2}(1)}{n}$. Expanding it with the same procedure as that used in the proof of theorem 1, one is able to find the estimator in fact is a partial pairwise estimator that focuses on the co-observations between $\{x|x \in \Sigma_{k1} \wedge \#(x) = 2\}$ and $\{y|y \in \Sigma_{k2} \wedge \#(y) = 2\}$. In addition, we can selectively pair data with model as that used in [19] to eliminate a local estimator sharing its member with other local estimators. Further, some of the estimators in (16) can be combined together as an estimator. In fact, (4) is a composite likelihood estimator that takes into account of all possible co-observations in Σ_k and gives the same weight to each of them.

C. Robustness

As theorem 1.2) points out, the estimate obtained by (4) is correct if and only if $\forall x, I_k(x) > 0, x \in \Sigma_k$. If some of the $I_k(x) = 0$, a different likelihood equation is needed in order to have a correct estimate. [22] classifies the data sets into five classes. Besides (4), another four likelihood equations are proposed, one for a data set that has not been considered previously. However, that is only a small portion of the cases that require new estimators. To handle them, we need to check the data set provided for estimation, and then select an estimator that fits to the data set. However, this issue has never been raised previously because

- there is a lack of other alternatives except full likelihood estimations, and
- there is a lack of knowledge about the correspondence between data and model.

Given the analysis and the estimators presented in this paper, checking and selecting become not only possible but also feasible. If the data required by $A_k(i)$ exist, the estimate obtained by the estimator is considered correct. In this regard, the estimators proposed in this paper can be viewed as trimmed estimators since each of them only requires a part of the information in a data set. To rank the robustness of the estimators proposed in this paper, we have

$$rank(A_k(i)) \geq rank(A_k(j)), \text{if } 2 \leq i < j \leq |d_k|$$

since if $A_k(i)$ to be invalid, there must have $\exists x, I_k(x) = 0 \wedge \#(x) = i$. If so, all other $A_k(j), j \in \{i+1, \cdot\cdot, d_k\}$ must be invalid as well since there must have $\forall y, I_k(y) = 0, x \subseteq y, y \in \Sigma_k$. To handle all cases that have $\exists x, I_k(x) = 0, x \in \Sigma_k \wedge \#(x) \geq 2$, we can either select an estimator from (16) that is not related to $I_k(x)$ or remove $I_k(x)$ and $\prod_{j \in x} \gamma_j$ from $A_k(\#(x))$.

D. With Missing Data

Loss rate estimation with a part of data missing has been considered in [23], where an expectation maximization procedure is used to approximate an estimate that corresponds to the full likelihood. With composite likelihood, this problem can be handled differently without modeling the missing data process as the method proposed in [24]. We can either

- select an estimator that is not related to missing data as discussed in VI-C, or
- add weight parameters to the likelihood function proposed in (3), where the likelihood objects involving missing data have a weight corresponding to the amount of missing for the models of missing at random (MAR) and missing completely at random (MCAR).

Then, an explicit estimator can be obtained by using the same method proposed in Section IV. The weight assigned to a likelihood object is inversely proportional to the amount of missing data.

VII. Conclusion

This paper aims at finding inspirations that can lead to simple and accurate estimators for loss tomography. To achieve the goal, a well known full likelihood estimator previously proposed is analyzed on the basis of the necessity of using full likelihood. The results obtained from the analysis show there are alternatives to connect data to models instead of using full likelihood. Then, the sufficient statistics identified previously for the full likelihood model are divided into a number of subsets according to a σ-algebra. Each of the subsets consists of a set of sufficient statistics for a model. Linking a subset to the model generating the subset leads to an estimator that measures the fitness between the two and can be solved explicitly. In light of this, a deep investigation has been carried out that leads us to the composite likelihood that has drawn considerable attention in recent years to estimate unknown parameters relating complicated correlations. Using

the composite likelihood, a set of composite likelihood functions are proposed according to the correspondences between data and models, and subsequently a set of explicit estimators are put forward. In addition, the properties of the estimators are investigated and presented in the paper that show the accuracy of an estimate is proportional to the number of descendants. The explicit estimators turn the headache that troubles researchers for many years into an asset. Apart from presenting the statistical properties in lemmas and theorems, a series of simulations are conducted and the results are reported that show the estimators proposed in this paper perform as good as the full likelihood estimator. The strategy and method used in this paper can be extended to the general topology.

REFERENCES

[1] R. Cáceres, N.G. Duffield, J. Horowitz, and D. Towsley. Multicast-based inference of network-internal loss characteristics. *IEEE Trans. on Information Theory*, 45, 1999.

[2] R. Cáceres, N.G. Duffield, S.B. Moon, and D. Towsley. Inference of Internal Loss Rates in the MBone . In *IEEE/ISOC Global Internet'99*, 1999.

[3] R. Cáceres, N.G. Duffield, S.B. Moon, and D. Towsley. Inferring link-level performance from end-to-end multicast measurements. Technical report, University of Massachusetts, 1999.

[4] M. Coates and R. Nowak. Unicast network tomography using EM algorthms. Technical Report TR-0004, Rice University, September 2000.

[5] B. Xi, G. Nichailidis, and V.N. Nair. Estimating netwrok loss rates using active tomography. *JASA*, 2006.

[6] T. Bu, N. Duffield, F.L. Presti, and D. Towsley. Network tomography on General Topologies. In *SIGCOMM 2002*, 2002.

[7] V. Arya, N.G. Duffield, and D. Veitch. Multicast inference of temporal loss charateristics. *Performance Evaluation*, 9-12, 2007.

[8] N.G. Duffield, J. Horowitz, F. Lo Presti, and D. Towsley. Explicit loss inference in multicast tomography. *IEEE trans. on Information Theory*, 52(8), 2006.

[9] W. Zhu and Z. Geng. Bottom up inference of loss rate. *Journal of Computer Communications*, 28(4), 2005.

[10] D. Guo and X. Wang. Bayesian inference of network loss and delay characteristics with applications to tcp performance predication. *IEEE trans. on Signal Processing*, 51(8), 2003.

[11] W. Zucchini. An introduction to model selection. *Journal of Mathematical Psychology*, 44, 2000.

[12] W. Zhu. An efficient loss rate estimator in multicast tomography and its validity. In *IEEE International Conference on Communicaation and Software*, 2011.

[13] W. Zhu and K. Deng. Loss tomography from tree topologies to general topologies. *arXiv:1105.0054*, 2011.

[14] J.E. Besag. Spatial interaction and the statistical analysis of lattice system (with discussion). *Journal of Royal Statistical Society*, 36, 1974.

[15] C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3), 2005.

[16] C. Varin. On composite marginal likelihoods. *ASTA Advances in Statistical Analysis*, 92(1), 2008.

[17] X. Xu and R. Reid. On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference*, 2011.

[18] Z. Jin. *Aspects of Composite Likelihood Inference*. PhD thesis, University of Toronto, 2010.

[19] G. Liang and B. Yu. Maximum pseudo likelihood estimation in network tomography. *IEEE trans. on Signal Processing*, 51(8), 2003.

[20] L. A. Goodman. On the Exact Variance of Product. *Journal of the American Statistical Association*, 55(292), 1960.

[21] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 1974.

[22] W. Zhu. Fitting a model to data in loss tomography. In *arXiv:1107.3879*, 2011.

[23] N.G. Duffield, J. Horowitz, D. Towsley, W. Wei, and T. Friedman. Multicast-based loss inference with missing data. *IEEE Journal on Selected Area in Communication*, 20(4), 2002.

[24] G.Y. Yi, L. Zeng, and R.J. Cook. A robust pairwise likelihood method for incomplete longitudinal binary data arising in cluster. *The Canadian Journal of Statistics*, 39(1), 2011.